Early Predictability of Asylum Court Decisions

Matt Dunn
Center for Data Science
New York University
mtd368@nyu.edu

Levent Sagun
Courant Institute
New York University
sagun@cims.nyu.edu

Hale Sirin
Humanities Center
Johns Hopkins University
hsirin1@jhu.edu

Daniel Chen
Toulouse School of Economics
Institute for Advanced Study
daniel.chen@iast.fr

Abstract

In the United States, foreign nationals who fear persecution in their home country can apply for asylum under the Refugee Act of 1980. Over the past decade, legal scholarship has uncovered significant disparities in asylum adjudication by judge, by region of the United States in which the application is filed, and by the applicant's nationality. These disparities raise concerns about whether applicants are receiving equal treatment under the law. Using machine learning to predict judges' decisions, we document another concern that may violate our notions of justice: we are able to predict the final outcome of a case with 80% accuracy at the time the case opens using only information on the identity of the judge handling the case and the applicant's nationality. Moreover, there is significant variation in the degree of predictability of judges at the time the case is assigned to a judge. We show that highly predictable judges tend to hold fewer hearing sessions before making their decision, which raises the possibility that early predictability is due to judges deciding based on snap or predetermined judgments rather than taking into account the specifics of each case. Early prediction of a case with 80% accuracy could assist asylum seekers in their applications.

1 Introduction

In the United States, foreign nationals who "demonstrate that they were persecuted or fear persecution due to race, religion, nationality, political opinion, or membership in a particular social group" (www.uscis.gov, 2016) can apply for asylum under the Refugee Act of 1980 (as illustrated in Figure 1), in compliance with international law, specifically the United Nations Protocol relating to the Status of Refugees of 1968 (Rottman et al., 2009). Asylum officers, immigration judges, members of the Board of Immigration Appeals, and judges of U.S. courts of appeals render approximately 79,000 asylum decisions each year (Rottman et al., 2009).

Over the past decade, legal scholarship has uncovered significant disparities in asylum adjudication by judge, by the US region in which the application was filed, and by the nationality of the applicant Ramji-Nogales et al. (2007). These differences are fundamentally at odds with the principle that all cases should receive equal treatment before the law. Similar cases should have similar outcomes in order to be fair. Perhaps even more importantly, such consistency is desirable because it demonstrates that the adjudications determining the asylum seeker's future do not depend on the personal opinions and prejudices of the individual judges to which the case happens to be assigned. To investigate the degree of consistency in court decisions related to similar asylum cases Ramji-Nogales et al. (2007) demonstrated that there is dramatic variation (Kleinberg et al., 2017) in decision-making among

different offices, regions, and officials, stating that "the variation is particularly striking when one controls for both the nationality and current area of residence of applicants, and examines the asylum grant rates of the different asylum officers who work in the same regional building, or immigration judges who sit in adjacent courtrooms of the same immigration court" (Rottman et al., 2009) (p.302).

A number of factors could be causing the dramatic differences in grant rates between judges. These include judge 'burnout', the theory that "the overwhelming caseloads and long hours worked without overtime...can potentially affect the outcome for applicants whose fates rest in judges' hands" Lustig et al. (2008). Marouf (2010) have suggested that immigration judges' work environments produce implicit bias that can drive their decision-making. For further related work, see Chen (2017) which explores the effects of NFL football games and weather on immigration decisions, and Chen and Eagel (2017) which predicts decisions focusing on all the information available to the statistician at the time the case closes. Anecdotal evidence also suggests that there are simply a few 'bad apples' – a subset of immigration judges whose decisions are deemed to be unfair toward the applicant Legomsky (2010).

This paper makes a conceptual distinction between inter-judge disparities in *predictions* and inter-judge disparities in *prediction accuracy*. A prediction refers to the model's estimation of whether a judge will grant or deny an asylum application, while prediction accuracy is the correlation between the model's predictions of judges' decisions and the actual results. Previously documented inter-judge disparities in predictions are inconsistent with equal predictions of the outcomes of cases before an applicant comes to court based solely on her case facts. Inter-judge disparities in prediction accuracy raises a different question. If case outcomes could be completely predicted after a particular judge is assigned, but prior to judicial inquiry into the case, this would indicate that judges did not take into account any non-coded differences between cases. To be sure, there may be cases for which the country and specific date of application should completely determine outcomes, for example in the case of a large-scale violent conflict in a particular country. However, significant inter-judge disparities in predictability suggest that this understanding of the country circumstances does not apply to all judges.

This current study focuses on predicting whether asylum is granted or denied based on the common features of a given asylum case: nationality, language, notice to appear (NTA), base city, hearing location, case type, attorney, and judge. The goal is to provide better information to asylum seekers regarding the strength of their application at the point when they are scheduled to appear before an immigration judge. The asylum seeker may believe that her case rests primarily on the specifics of her story and may not be aware of the extent of the importance of external factors, such as the judge to whom her case is assigned. Therefore, we developed a predictive model to help applicants understand how these external factors might affect their application.

Our model allows an asylum applicant to predict the final outcome of her application with 80% accuracy. The model uses all features of the application available at the time the applicant receives an NTA. Our model also allows us to evaluate the relative impact of specific features. Echoing the findings reviewed above, the features that have the strongest impact on an application's final outcome are the adjudicating judge and the nationality of the applicant. In addition, although cases are randomly assigned to the judges (Ramji-Nogales et al., 2007), this study shows that a certain percentage of the judges are highly predictable, and almost always either grant or reject asylum applications regardless of the specifics of the case at hand. This suggests that personal predilections can be a major factor in judges' decisions.

We also make a conceptual distinction between predictability and early predictability. We compare *early* predictability (using only information available at the time the case opens) with predictability (using all the information available in our data at the time the case closes). This comparison raises questions about judges' use of snap judgments, heuristics, and predetermined judgments when deciding cases (Ambady and Rosenthal, 1993). Stereotypes have been found to influence impression formation in courtrooms and tend to be more pronounced under conditions that generally foster heuristic processing, such as time pressure and distraction (for a review, see Bless et al. (1996)), factors that have been raised in regards to the asylum courts.

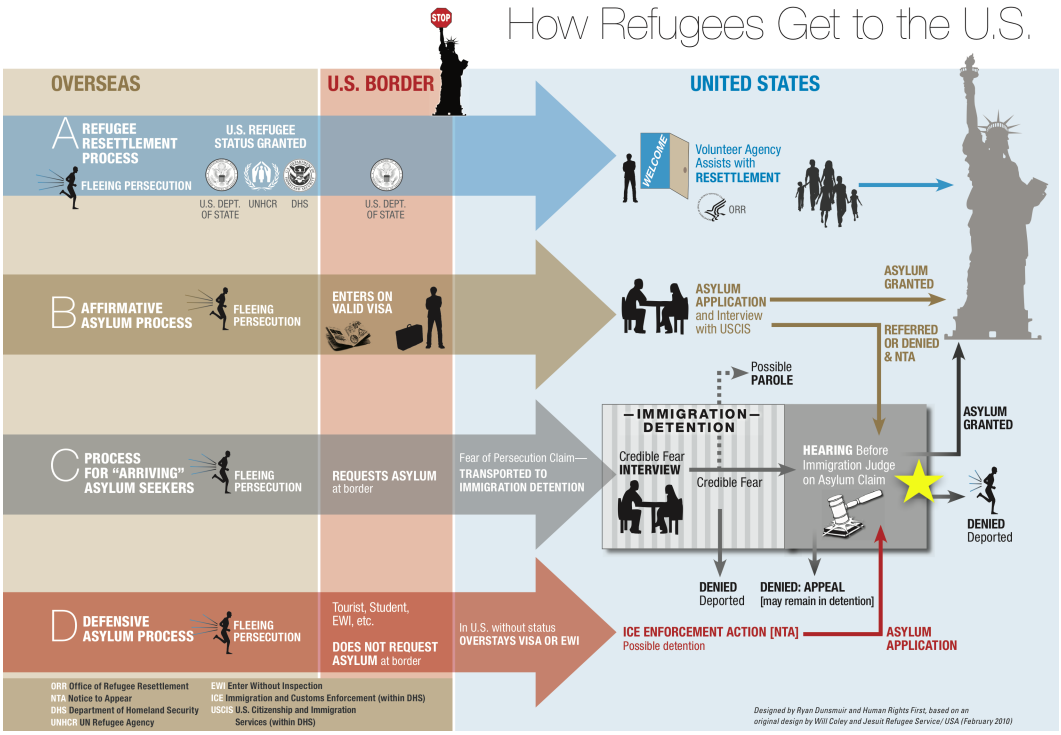


Figure 1: The asylum application process illustrated. Source: `www.rcusa.org` (2016). Note that our model attempts to predict the immigration judge's decision (the starred branch in the process diagram).

2 Data and Methodology

2.1 Summary Statistics

In the course of sorting the data to build our model, we found several statistics worthy of note. Table 1 shows the number of cases per state. We see that the vast majority of asylum applications are filed in three states, New York, California, and Florida. Table 3 demonstrates that there is substantial variation in average grant rates when citizens of different countries apply for refugee status in the United States.

The model is built using data from the Executive Office for Immigration Review (EOIR), which issues NTAs. A pre-processed version of the same data was used in Chen et al. (2016), which explores the gambler's fallacy in immigration decisions. The current article, however, focuses on predicting the outcomes of asylum applications at the time the case opens using machine learning techniques. This approach addresses both the behavioral question of the role of snap judgments (Ambady and Rosenthal, 1993) and the policy question of how best to advise applicants when they receive an NTA. The raw data include multiple records for a given case. Further details on the EOIR data include:

- Data on the scheduled time and outcome of hearings: 15,377,520 records, which include approximately 70 additional features about the hearings, such as location and the presence/absence of an attorney.
- Data on the outcome of asylum cases: 6,084,435 records, including information on the nationality of applicant, case type, asylum seeker type, base city, hearing location, decision type, attorney present/absent, and unique judge identifier.
- Data on the biography of the judges: 455 text files with paragraph biographies of the judges.

The project is based on two pre-processed views of these data: (1) a merged view of the raw courts data, in which each record corresponds to an asylum application decision and (2) a hand-coded set of judicial biographical features. Once the data are merged we have:

Table 1: Ten states with the highest number of asylum applications.

State	Count	Percentage	Grant Rate
NY	155,656	0.26	0.51
CA	134,697	0.22	0.32
FL	100,368	0.17	0.23
TX	29,674	0.049	0.26
NJ	22,893	0.038	0.35
MD	19,291	0.032	0.4
VA	17,680	0.029	0.39
MA	14,598	0.024	0.32
IL	13,179	0.022	0.4
PA	12,621	0.021	0.32

- **Total pre-processed data set**: 602,500 cases
- **Total number of grants**: 213,731 cases (35.5%)
- **Training/Test split**: 482,000/120,500 cases

2.2 Data Engineering

Our objective was to build a model that could predict whether an applicant would be granted asylum at the time he was notified about his initial hearing time, location, and judge assignment. In order to avoid data leakage, we first constructed a data dictionary defining each feature and indicating whether it would be available at the time this initial notification. We based our data dictionary on information from Ramji-Nogales et al. (2007), as well as conversations with practicing immigration attorneys. The final dataset includes: application decision (target as a binary variable), language (spoken by the applicant), nationality (of the applicant), base city (asylum seeker is assigned to one of several regional immigration courts), attorney (whether the asylum seeker was represented by an attorney), date of the NTA, hearing location (e.g. regional courthouse, detention center), case type (affirmative and defensive)[1] judge features (unique identifiers for individual judges in addition to general features about the judge, such as gender and work history).[2]

2.3 Model selection

Two main characteristics of the data led us to use random forest algorithms: (1) the dominance of categorical features and (2) collinearity when restricted to our baseline models. Most of the features are categorical variables, with categories that have hundreds of possible values. A common way to featurize this is to declare each variable value as a feature and tag them using binary variables. However, this dramatically expands the dimension of the feature space that makes the data susceptible to memory issues. While assigning numerical values to categories keeps the feature space small, the lack of meaningful scale in the features rules out nearest-neighbor classifiers. The data used for early prediction is linearly separable only when one includes the date- and time-specific variables that are not inherently binding to the case. Without those, many applicants have identical features but different decisions: even in the full model with 32 features that include year and month information, there are 72,853 case types. This rules out linear classifiers, and even some mildly nonlinear SVMs

[1] Affirmative applications are made by the asylum seeker voluntarily, within one year of arriving in the United States, and are not triggered by a removal order from the U.S. government (`www.uscis.gov`, 2015). A defensive application is one in which the asylum seeker has requested asylum to prevent their removal from the United States. The case for a defensive application is presented in front of an immigration judge in adversarial proceedings and is subject to cross-examination by a government attorney (Rottman et al., 2009). The U.S. government does not provide a defense attorney; at her own expense, or with the support of a non-profit advocacy group, the applicant can request that her case be presented by an attorney (Rottman et al., 2009). See Figure 1 for a description.

[2] Ramji-Nogales et al. (2007) also uncovered a number of significant predictors of grant rates, including the gender and employment history of the immigration judge, as well as significant regional differences in grant patterns between asylum courts.

(support vector machines). Nor are neural networks suitable for the task at hand, for the two reasons described above.

Since our preliminary results indicate no significant improvement over random guessing when the data are trained with SVMs and neural networks, we analyzed the data using random forest algorithms. Random forests work well in the context of such data limitations. Tree algorithms are powerful in handling categorical variables due to their inherent branching structure. Moreover, ensemble methods combined with randomness can overcome the ambiguity created by same-input-different-output cases.

3 Results

Based on our discussions with practicing lawyers, we suspected that the identity of the presiding judge would be a significant factor in determining whether an applicant's application would be granted or denied. To test this belief, and to understand how an applicant's characteristics (nationality and language) and case information (case type and application type) impacted the random forest model, we incrementally added features by training the model on each set of features and then testing its performance on a hold-out test set. We first trained the model on the smallest reasonable feature space, and then added features and trained the model on each set of data as the feature space became more complex. This approach revealed how attributes of the applicant and judge affected the accuracy of the model.

3.1 Incremental Evaluation of Feature Space

Our dataset includes 602,500 records, of which 35% of the cases were granted asylum. With 80/20 splitting, we have 482,000 rows in the training set and 120,500 in the test set, and the percentage of applications that were granted is 35% for each. For all the models shown in Table 2, we used a grid search of the random forest algorithm with $\{128, 256, 512, 1024\}$ trees and sevenfold cross validation over the training set. This table reports the area under the receiver operating characteristic curve (AUC). The receiver operating characteristic curve (ROC) plots the true positive rate of a classifier against the false positive rate. A naive classifier would generate an AUC of 0.5 and a perfect classifier would generate an AUC of 1.0.

Table 2: Accuracy and ROC AUC across models.

Model	Accuracy	ROC AUC
Part 1	0.71453	0.74101
Part 2	0.76441	0.82056
Part 3	0.73219	0.77484
Part 4	0.77816	0.83964
Full Model	0.81589	0.88137

Baseline Models:

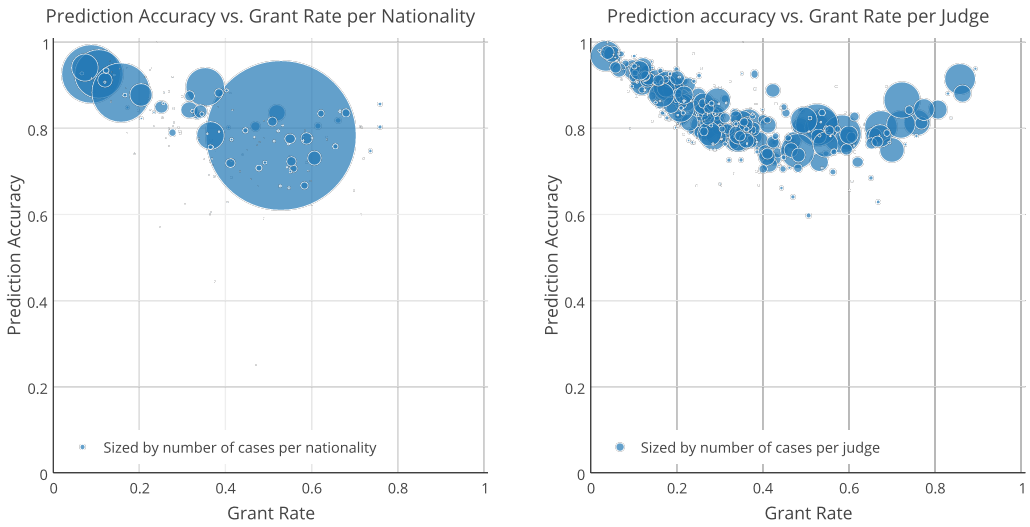
(Part 1) Judge ID – To test whether the presiding judge was a significant factor in predicting the outcome of an asylum application, our baseline model used a single feature, the judge ID. With this single feature, the trained random forest model was able to predict whether someone would be granted or denied asylum with a mean accuracy score of 71%.

(Part 2) Judge ID and Nationality – Adding the applicant's nationality results in a major jump of 5% in accuracy, and an even more significant jump in the ROC AUC score.

(Part 3) Judge ID and NTA – When we replace nationality with NTA date and train the model, we see a minor improvement in the prediction accuracy of 2%, which is less effective than **Part 2**.

(Part 4) Judge ID, Nationality, and NTA – Comparison with **Part 2** and **Part 3** suggests that variation over time appears to have little additional impact on the outcome of asylum decisions. In other words, the asylum decision patterns regarding nationality did not significantly change over the 20-year period.

This is rather striking, since we have limited ourselves to only the most basic information about the applicant. One interpretation of this finding is that a large component of the final decision regarding



(a) Grouped by applicant's nationality. For more information on large dots, please refer to Table 3.

(b) Grouped by individual judges (cases are randomly assigned).

Figure 2: The x-axis shows the variance in grant rates. The y-axis shows the accuracy of our model on cases when they are grouped according to the given class.

the asylum application is already set prior to the judge's review of the application. Additional review of the case by the judge may have less impact than may be desired on the application's outcome. This brings us to the full model.

Full model at the time the case closes: For our final model, we included all the features noted in Section 2.2. The final item, "Judge Features" (such as gender, law school, bar, and active president), was obtained from another dataset; merging the two trimmed the size of the dataset by about 20%, keeping the overall statistics the same. The full model yields a mean accuracy score of over 80% on the hold-out test dataset.

3.2 Discussion

Using only the unique identifier of the presiding judge, we are able to predict with 71% accuracy whether an applicant will be granted or denied asylum in the United States. We can see from Figure 2a that when we train the random forest model on the complete feature space, nationality is still a primary driver of whether or not an individual applicant will be granted asylum. This is demonstrated by the fact that the average grant rate varies substantially even though there is little variation in the predictive accuracy. This calls into question how much of the individual application's facts or litigation strategy is materially relevant to its final outcome.

Including all the features available on the date the case opens, the predictive accuracy is 80%. Additional variation in predictive accuracy appears across judges, even holding fixed their grant rate per nationality, as seen in Figure 2b. This figure shows that some judges are fairly conservative, granting asylum to less than half of applicants, and as their grant rate falls, prediction accuracy increases. Some of this may be mechanical: in the extreme, a judge who never (or always) grants asylum will be very predictable. Some of the variation in the y-axis dimension can shed light on whether some judges vary in their attention to applicants from certain nationalities. Alternatively, applicants from certain nationalities may be very homogeneous, but homogeneity in unobserved applicant characteristics is unlikely to explain the variation in predictability across judges since cases are randomly assigned.

4 Conclusion

While Ramji-Nogales et al. (2007) noted regional differences in asylum grant patterns and identified several judge-level variables correlated with asylum grant rates, they did not explicitly aim to build

Table 3: Top ten countries by the number of applications and their grant rate.

Country	Count	Percentage	Grant Rate
China	107964	0.19	0.53
Haiti	42013	0.074	0.16
El Salvador	41626	0.074	0.087
Guatemala	34705	0.061	0.11
Colombia	27713	0.049	0.35
India	19161	0.034	0.37
Mexico	19031	0.034	0.073
Nicaragua	15987	0.028	0.2
Albania	12036	0.021	0.52
Indonesia	11399	0.02	0.32

predictive models based on these features. Thus, we sought to extend their work by building a predictive model. Our goal was to develop a model that could predict whether an applicant would be granted asylum, using only information available at the time an applicant receives an NTA. In the process, we also aimed to interpret what early predictability might mean.

University and pro-bono asylum law clinics have limited resources and a large number of prospective clients. A predictive model could potentially assist these organizations by allowing them to estimate the probability that an applicant will receive asylum prior to any case assistance. This could also allow them to suggest interventions that improve the odds of success.

Individual judges could also use the model as a feedback tool to gain insight into their past granting patterns. A judge could thus enter basic information about the case and her own ID, and receive a probabilistic prediction of her decision on the application, based on past behavior.

4.1 Model Deployment Considerations

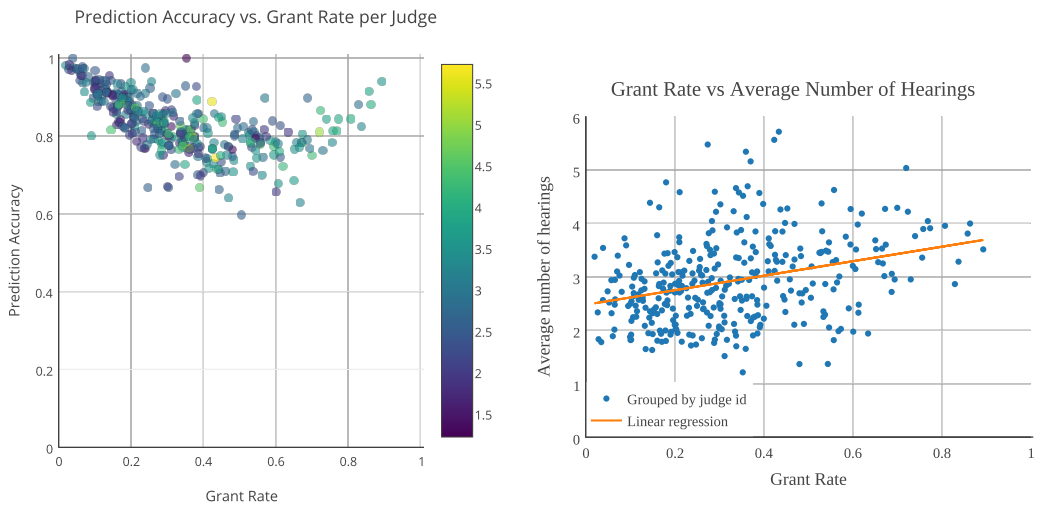
Finally, while our model achieved a high test set AUC, prior to deployment we would suggest validating the model against a recent test set. Given the known non-stationarity of the asylum process (often attributed to domestic political transitions), it would be necessary to assess performance under the current interpretation and enforcement of asylum law.

Moreover, incorporating international or domestic policy influences could be a further area of investigation to help improve this predictive model. In Salehyan and Rosenblum (2008) the authors examine which domestic and foreign policy factors affect asylum decisions. They show that while humanitarian conditions are a consistently important predictor of such decisions, applicants were significantly less likely to be granted asylum if they fled countries that were U.S. military allies in the 1980s, U.S. trade partners in the 1990s, or important sources of undocumented migration in the 1980s and 1990s (Salehyan and Rosenblum, 2008).

To increase the utility of the predictive model, an additional area of work would be to develop another (simplified) model to determine which asylum court offers an asylum seeker the highest estimated probability of being granted asylum. Given the anticipated use of our model (to estimate an applicant's probability of success when they receive an NTA), it would be beneficial to develop supporting models that highlight the potential margins that the applicant could exploit to increase his or her chances of success in court. Moving to the catchment area for a different asylum court represents one of few such margins (beyond seeking legal representation). A supporting model could indicate the anticipated effect of such a move.

4.2 Early Predictability

This paper raises the concern that judges might not be incorporating sufficient case information when deciding asylum applications, which raises a separate question from the significant inter-judge disparities documented by prior work. We have found that judges who are highly predictable tend to hold fewer hearings, on average, before rendering their decision, as shown in Figure 3a, which may indicate that highly predictable judges are employing snap or predetermined judgments. It also



(a) Prediction accuracy vs. average grant rate, looking at the average number of hearings per case per judge.

(b) Grouped by individual judges (cases are randomly assigned).

Figure 3: Average number of hearings vs. grant rate.

suggests that less predictable judges are not simply flipping a coin, because they are holding more hearing sessions.

The early predictability we document at the time the case opens relative to when it closes might be due to the court database not recording information that is relevant to explaining the final decision, which could suggest that court databases should be recording additional information. Yet this possibility does not diminish the relevance of our inter-judge findings, because we find judges who hold more hearing sessions to be less predictable, suggesting that the hearing session data has some informational content.

We also found that judges who tend to grant asylum more often hold more hearing sessions per applicant than those who almost always deny asylum, as shown in Figure 3b. This is not a mechanical result, and it raises further questions regarding which sets of judges are fairer about taking into account the individual differences between cases rather than exercising snap or predetermined judgments. To be sure, it is possible that judges who use fewer hearing sessions are trying to be more sparing of the court's time and resources and, as a consequence, become more predictable. Whether early predictability is due to judges not considering applicant circumstances or simply being constrained on time and resources, using this study to inform more systematic reform along the lines of Legomsky (2010) is a final—and more ambitious—possibility.

References

Nalini Ambady and Robert Rosenthal. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology* 64, 3 (1993), 431.

Herbert Bless, Norbert Schwarz, and Markus Kemmelmeier. 1996. Mood and stereotyping: Affective states and the use of general knowledge structures. *European review of social psychology* 7, 1 (1996), 63–93.

Daniel Chen. 2017. *Mood and the Malleability of Moral Reasoning*. Technical Report. National Bureau of Economic Research.

Daniel Chen and Jess Eagel. 2017. *Can Machine Learning Help Predict the Outcome of Asylum Adjudications?* Technical Report. National Bureau of Economic Research.

Daniel L. Chen, Tobias J. Moskowitz, and Kelly Shue. 2016. Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires. *The Quarterly Journal of Economics* 131, 3 (2016), 1181. DOI:http://dx.doi.org/10.1093/qje/qjw017

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. *Human Decisions and Machine Predictions*. Working Paper 23180. National Bureau of Economic Research. DOI:http://dx.doi.org/10.3386/w23180

Stephen H Legomsky. 2010. Restructuring Immigration Adjudication. *Duke Law Journal* (2010), 1635–1721.

Stuart L Lustig, Niranijan Karnik, Kevin Delucchi, and Lakshika Tennakoon. 2008. Inside the judges' chambers: Narrative responses from the National Association of Immigration Judges stress and burnout survey. *Geo. Immigr. LJ* 23 (2008), 57.

Fatma E Marouf. 2010. Implicit Bias and Immigration Courts. *New Eng. L. Rev.* 45 (2010), 417.

Jaya Ramji-Nogales, Andrew I Schoenholtz, and Philip G Schrag. 2007. Refugee roulette: Disparities in asylum adjudication. *Stanford Law Review* (2007), 295–411.

Andy J Rottman, Christopher J Fariss, and Steven C Poe. 2009. The Path to Asylum in the US and the Determinants for Who gets in and Why. *International Migration Review* 43, 1 (2009), 3–34.

Idean Salehyan and Marc R Rosenblum. 2008. International relations, domestic politics, and asylum admissions in the United States. *Political Research Quarterly* 61, 1 (2008), 104–121.

www.rcusa.org. 2016. Refugee Council USA. Asylum and Detention. http://www.rcusa.org/asylum-and-detention. (2016).

www.uscis.gov. 2015. Obtaining Asylum in the United States. https://www.uscis.gov/humanitarian/refugees-asylum/asylum/obtaining-asylum-united-states. (2015).

www.uscis.gov. 2016. US Citizenship and Immigration Services. Refugee definition. https://www.uscis.gov/humanitarian/refugees-asylum/refugees. (2016).